Preregistration FLO paper

Eva Viviani

Michael Ramscar

Elizabeth Wonnacott

15 June 2020

# Study information

In this preregistration we describe a detailed plan to replicate Ramscar et al. (2010) paper about the effect of the order of presentation of feature-label pair association on symbolic learning. We plan to translate the experiment in an online setting and to introduce a contingency judgment task that would link the theoretical predictions about error-driven learning, i.e., cue-competition, to the participants’ ratings behaviour of the plausibility of the feature-label association.

We expect that this commitment will provide (1) a robust paradigm that can be used as starting point for future studies online, (2) a sensitive task capturing estimates of cue-competition across participants, providing a direct comparison to the predictions based on the RW model (Rescorla, Wagner, and others (1972)). We base our predictions on data of an unpublished study that replicates Ramscar et al. (2010) effects and on our own pilot data.

# Research Questions

Does Feature-Label-Ordering training affect discrimination learning?

* 1. Does frequency of occurrence of the predictive cues affect learning?
  2. Does feature-label and label-feature order affects learning?
  3. Do feature-label and label-feature order groups differ when the frequency of presentation is low?

# Hypotheses

We translate our hypothesis in terms of average accuracy estimated in two tests previously implemented by Ramscar et al. (2010), label-picture and picture-label tasks. These tasks require participants to select the right label-feature association among four alternatives.

We expect:

1. **A main effect of frequency:** the association between predictive cue and its correct label in the low frequency condition will be extremely challenging compared to the high frequency condition regardless of the learning. Therefore we expect higher accuracy in the high vs low frequency condition. From the study that replicates Ramscar et al. (2010), [ = 1.7, SE = 0.24, z value = 6.85, *p* 0.0001].
2. **A main effect of learning:** the association between predictive cue and its correct outcome will be effectively learned only if the training enhances discriminative learning, that is, only in the feature-label training. Therefore we predict higher accuracy in the feature-label learning, compared to label feature learning. From the study that replicates Ramscar et al. (2010), [ = 0.58, SE = 0.14, z value = 4.12, *p* 0.0001].
3. **An interaction between frequency and learning:** we predict higher accuracy in the low frequency condition for the feature-label learning group compared to label-feature. From the study that replicates Ramscar et al. (2010), [ = 1.02, SE = 0.26, z value = 3.92, *p* = 0.0001].

We will also attempt to measure cue competition in a contingency judgment task, for potential sources of comparisons inherent to the RW computational model (Rescorla, Wagner, and others (1972)). In this task participants are required to express their confidence regarding the label-feature association presented to them. Although this is still an exploratory test, we base our hypothesis on a pilot previously conducted in the lab and on the estimates of the computational model.

For what it concerns a 100% match between label and the fribble presented (match scenario), we expect only positive ratings because the cue-outcome association is always positively reinforced.

However, we expect:

1. **A main effect of frequency:** Average ratings in the high frequency condition will be positive and higher than the low frequency condition. From our own pilot, [ = 47.17, SE = 10.6, z value = 4.44, *p* = 0.00004].
2. **An interaction between frequency and learning:** average ratings in the high frequency condition will be positive and higher for the label-feature group compared to the feature-label. From our own pilot, [ = 30.05, SE = 21.25, z value = 1.41, *p* = 0.16].

For what it concerns a mismatch between label and fribble presented where the high salience cue (body shape) is shared across labels in two different frequency conditions, i.e., blue body high frequency condition - label 1, blue body low frequency condition - label2 (mismatch-type1 scenario). We expect in this condition that cues will be downweighted because they do not predict outcomes in the 100% of the cases, however only if the learning promotes cue-competition.

Therefore, we predict:

1. **A main effect of frequency:** Average ratings in the high frequency condition will be negative and higher than the low frequency condition. From our own pilot, [ = -33.33, SE = 10.39, z value = -3.20, *p* = 0.002].
2. **An interaction between frequency and learning:** The downweighting of unreliable (non-discriminative) cues, is a peculiarity of the feature-label group only, and this will be evident in the low frequency condition. Therefore we predict that the average ratings in the low frequency condition will be positive for the label-feature group, while negative for the feature-label. From our own pilot, [ = 4.51, SE = 20.8, z value = 0.21, *p* = 0.82].

# Sampling plan

## Existing data

Registration prior to creation of data.

## Data collection procedures

* Participants will be tested via Gorilla (<https://gorilla.sc/>), and recruited through Prolific (<https://gorilla.co/>). Participants will complete independently the experiment online.
* All adults with consent to participate will be tested, however some exclusion criteria will apply for the analysis. In line with Ramscar et al. (2010), those that will score less or equal to 80% in the control condition will be removed from the analysis.
* Payment of £2 per session will be made to each adult participant (7£ per hour). Adult participants will be offered the full amount at the end of the session only.

## Sample size

* This study aim to test max 90 adults English native speakers. The sample size is informed by Nixon (2020) study (Experiment 2) with adults (93 participants), and Vujović, Ramscar, and Wonnacott (2019) (84 participants).

## Stopping rule

* Using Bayesian statistics, we are going to collect data until the Bayes factor is larger than 3. We will look at the data after 25 participants in each Learning Group (feature-label/label-feature). If the Bayes factor is larger than 3, we will stop testing.

# Variables

## Manipulated variables

Learning:

* Label-to-Feature (LF) learning, i.e., learning to predict the objects from the labels.
* Feature-to-Label (FL) learning, i.e., learning to predict the labels from the objects.

Frequency:

* High frequency exemplars (highFreq): 75% of exemplars are made of a high salient feature specific of that subcategory.
* Low frequency exemplars (lowFreq): 25% of exemplars are drawn from another subcategory and have a different salient feature.

## Measured variables

Generalization tasks.

Accuracy on a speeded three-alternative forced choice test (4AFC) where participants:

* Match an unseen exemplar to the three category labels (56 trials)
* Match a label with three unseen exemplars (56 trials)

Contingency judgment task.

Judgement ratings where participants estimate the strength of the association between a label and an unseen exemplar. Trials will be 100% match between label and the exemplars (match, 24 trials), partial match between label and exemplars (mismatch-type1, 24 trials).

Half of the participants will match an unseen exemplars to the three category, and the other half will match a label with three novel exemplars. All participants will express their ratings as final task.

# Design Plan

## Study type

Experiment - A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This is also known as an intervention experiment and includes randomized controlled trials.

## Blinding

Participants will not know the learning group to which they have been assigned, neither the frequency manipulation.

## Study design

In the GLMMs (generalization tasks) and LMM (contingency judgment task), *frequency* and *learning* will be specified as fixed effects, and *(frequency|subjects)* as random factor. For the contingency judgment task, we are going to run two separate LM models for the type of trial (match and mismatch-type2). For the GLMMs we are going to specify a binomial distribution given the binary nature of the dependent variable (accuracy). For the Bayesian analyses we will be computing Bayes factors following the approach in Dienes (2008). We need a mean and SE to summarize the data and these will come from the estimates and standard errors of the relevant coefficients in the LMM and GLMMs. We inform H1 using data from a replication study of the Ramscar et al. (2010) paper, and our own pilot data for the contingency judgment task. We will specify a half normal distribution in the Bayes Factor computation.

## Randomization

Participants will be randomly allocated to one of the two learning conditions (feature-label or label-feature). The learning sequence presented is randomized for each participant and presented in two blocks. No other cover task is performed.

# Analysis Plan

## Statistical models

For the Bayesian analyses we will be computing Bayes factors using the method advocated by Dienes (2008). We will also alongside provide frequentist statistics. We work in log odds space to meet assumptions of normality and will model the H1 by using estimate of the mean for theory as the SD of a half normal (for one tailed) distribution.

Logistics Mixed Effects Models:

LMEs will be used to compute frequentist statistics in the picture-label and label-picture tasks. The choice of method is due to the nature of the dependent variable of correct (1) and incorrect (0) response. Frequency by participant effects are included in the mixed models. This is because frequency is expected to greatly vary across participants. Picture-label and label-picture task will be considered together in the analysis if no significant differences emerge across the two tasks.

Linear Mixed Effect models:

LMMs will be used to compute frequentist statistics in the contingency judgment task. The choice of method is due to the continuous nature of the dependent variable. Frequency by participant effects are included in the mixed models. This is because frequency is expected to greatly vary across participants. Match and mismatch-type1 trials will be considered in separate models.

We will then run pairwise comparisons with learning and frequency as predictors.

## Follow-up analyses

None.

## Inference criteria

We will base our inferences only on the Bayes analysis. We will continue to work in log odds space (as for Frequentist) to meet assumptions of normality, using estimates and standard errors which come from the logistic mixed effects models. Following Dienes (2008), we test one sided predictions, using estimates as the SD of a half normal distribution that will be derived from LMEs for intercept (effect of learning in each Condition) and interaction (main effect of Experiment and Age Group). Note that this means that the maximum we might expect is twice our estimate. In some cases we compute SD using knowledge of constraints on the likely maximum value. (Note- this approach differs from the approach of setting a uniform maximum, in that it favors smaller values, which seems more appropriate).

* The estimates for each key analyses are already given in the Hypothesis section, for each specific hypothesis listed.
* We will say we have substantial evidence for H1 if BF is larger than 3 and for H0 if BF is less than .33.

## Data exclusion

Participants that will score less or equal to 80% in the control condition will be excluded from the analysis.

## Missing data

If more than 10% of the responses at either test tasks from a participant are missing, that participant is excluded from the analyses.

## Exploratory analysis

We will run a correlation between the generalization tasks and the contingency judgment task.

# References

Dienes, Zoltan. 2008. *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. Macmillan International Higher Education.

Nixon, Jessie S. 2020. “Of Mice and Men: Speech Sound Acquisition as Discriminative Learning from Prediction Error, Not Just Statistical Tracking.” *Cognition* 197: 104081.

Ramscar, Michael, Daniel Yarlett, Melody Dye, Katie Denny, and Kirsten Thorpe. 2010. “The Effects of Feature-Label-Order and Their Implications for Symbolic Learning.” *Cognitive Science* 34 (6): 909–57.

Rescorla, Robert A, Allan R Wagner, and others. 1972. “A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement.” *Classical Conditioning II: Current Research and Theory* 2: 64–99.

Vujović, Maša, Michael Ramscar, and Elizabeth Wonnacott. 2019. “Vujović, Ramscar & Wonnacott (Pre-Print). Language Learning as Uncertainty Reduction: The Role of Prediction Error in Linguistic Generalization and Item Learning.”